Aiding Cancer Research by data analysis with the R-cran statistical Environment

Ahmad Barghash[a], Volkhard Helms[b], Sonja M. Kessler[c]
[a]School of Electrical Engineering and Information Technology, German Jordanian University, Amman, Jordan
[b]Center for Bioinformatics, Saarland University, Saarbruecken, Germany
[c]Department of Pharmacy, Pharmaceutical Biology, Saarland University, Saarbruecken, Germany

***Abstract*— The ability of generating large scale high-throughput biological datasets leads to major improvements in the analysis of cancerogenesis. Currently, one sample contains thousands of measurements in some datasets. Therefore, modern cancer analysis techniques start by the low-cost computational analysis of high-throughput datasets. The available datasets are often compiled in major public access databases like The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO). The emerging need to apply statistical approaches for analyzing large scale datasets led to the establishment of open source statistical environments such as (R-Cran).**

Keywords—R-Cran; Bioconductor; Cancer analysis; Colorectal cancer; PIK3CA;

I. INTRODUCTION

The genetic material of a living organism that is passed from one generation to the next one is packed in the genomic DNA. This is often called the book of life. The long genomic sequence of the DNA is based on a simple 4-character alphabet {A,C,G,T}. These characters refer to the nucleotide bases Adenine, Cytosine, Guanine, and Thymine, respectively. Besides the replication step, subsets of the DNA (genes) are constantly copied into messenger RNA by a process called transcription. The transcription of genes into messenger RNA is called gene expression. The messenger RNA sequences carry the needed codes to produce the proteins which perform a vast array of functions within the organisms.

Gene expression is sensitive to clinical conditions or toxic agents and can be used as a marker for certain biological processes because of its clear up/down regulation states. Formerly, analysis of gene expression was conducted in a low-throughput fashion where one gene is analyzed at a time (Northern blotting). The invention of the microarrays in the 1990s enabled researchers to evaluate the expression of many genes at once. Nowadays, thousands of genes can be analyzed on one microarray chip. Alternatively, gene expression is measured by the recently introduced RNAseq technology that also provides access to single nucleotide variants of the expressed sample.

Regardless of the technology used, gene expression arrays are presented as an array (n x m) of numerical values where n and m correspond to the number of genes and experiments, respectively.

Gene expression analysis is widely used in the characterization of diseases [1] [2] [3] [4], the diagnosis of possible drug targets [5], and in analysis of gene networks or gene-gene interactions [6]. Additionally, it plays a key role in modern cancer analysis. It is often needed to check whether a gene shows differential expression behavior in tumor samples compared to normal samples. Several statistical approaches are often used to analyze the gene expression datasets.

Recently, many statistical environments that can aid in gene expression analysis became available to researchers. Such environments include the widely spread packages matlab, SPSS, Octave, and R-Cran. Wheras matlab and SPSS are commercial softwares, Octave and R-cran are available for free. However, octave has fewer libraries compared to R which has around 6000 libraries. For example, Bioconductor is an open source software that includes many tools for the analysis and interpretation of high-throughput genomic data. Bioconductor is based primarily on the R programming language and runs in its environment.

In this work, we analyze expression data downloaded from public sources in the R-cran statistical environment to identify or validate possible tumor marker genes. The downloaded data from the public sources like TCGA and GEO [7] often contain the raw numeric values describing the gene expression. Such datasets needs to be processed by additional normalization and mapping steps before the biological analysis can be started. This can be performed using several R packages such as affy [8] or limma [9].

Furthermore, expression datasets often suffer from outliers at the gene level [10]. Performing any analysis without searching for possible outliers might lead to inaccurate results. Recently, several outlier detection methods are introduced. Expression datasets should be first searched for outliers before performing the needed analysis. Outliers can be detected in R packages like parody [11].

The statistical tests often used to analyze differential gene expression includes the Kolmogorov-Smirnov (KS) test, the ordinary t-test, and the Welch's t-test. All tests compare different-sized samples with unequal variance. These tests are available in the basic R package.

978-1-5090-4580-8/16/$31.00 ©2016 IEEE

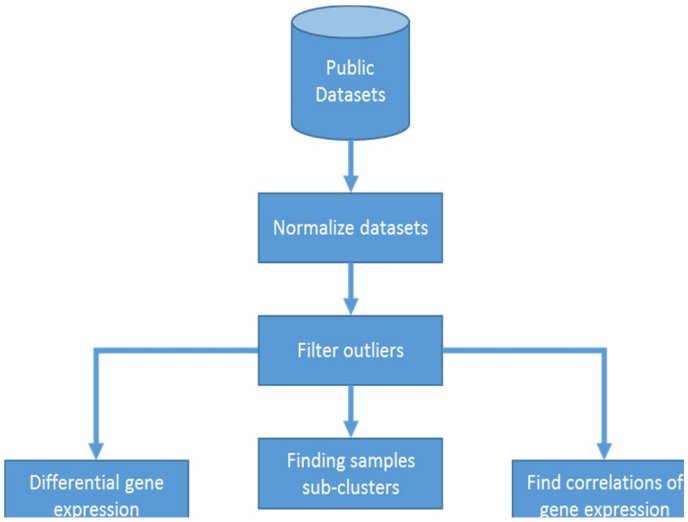

Fig. 1. Workflow for processing gene expression datasets and subsequent analysis

The gene expression datasets are searched for genes with similar expression behavior. Such genes might be co-regulated by the same set of transcription factors, may have similar functions or do participate in the same biological processes. The most common used statistical measure for detecting the similarity in expression profiles is the Pearson correlation coefficient (PCC). The PCC can be calculated using the basic or the Hmisc [12] R packages.

Additionally, gene expression datasets can be analyzed to identify specific expression patterns at the sample/experiment level (columns of the expression array). Clustering methods can be applied to the array samples to define sample sets with remarkable expression pattern and map them to certain disease conditions. Non-hierarchical clustering methods are less used in the analysis of expression datasets. However, all subtypes of hierarchical clustering (single, complete, average) are frequently applied to expression datasets. In this work, we apply hierarchical clustering only to tumor samples when searching for tumor sub-classes.

Hierarchical clustering can be performed using several R packages like cluster, fastcluster, in addition to the basic package. Once the clusters are determined, we test to which clusters the genes belong using the signal to noise ratio (SNR) described Hoshida [13].

II. Materials and Methods

A. The gene expression datasets analysis module

Fig. 1 presents a suggested analysis module for the expression datasets. The datasets (downloaded from TCGA or GEO [7] for example) might be in RAW format and thus needs further preparation and normalization using the affy [8] or limma [9] packages. Also, the normalized datasets might still suffer from outliers. Therefore, the datasets are tested for possible outliers using the parody [11] package for example. Once the dataset is ready, the major analysis starts.

B. The core analysis of gene expression datasets

The core analysis consists of three independent parts. The first part concentrates on identifying the differentially expressed genes. Depending on the gene expression data distribution, we select a suitable statistical test. If the data follows a normal distribution, then any type of t-test (Welch for example) is preferred (1). Otherwise, the KS test would detect the available differentially expressed genes.

$$t_j = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} \quad (1)$$

Where:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

s^2 is the unbiased estimator of the variance and n is the number of participants

The resulting p-values are used as evidence for strong differential expression and boxplots are generated accordingly from the R basic library.

In the second core part of the analysis, possible co-expressed genes are detected. The whole dataset is analyzed in R by measuring the level of co-expression between any pair of genes using PCC (2).

$$PCC = \frac{\sum_{i=1}^{n}(a_i - \bar{a}) \times (b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2} \times \sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} \quad (2)$$

Where:

a_i, b_j are the expression values of genes A and B respectively

$\bar{a}$, $\bar{b}$ represent the sample means for the genes

n is the number of samples

The final part of the analysis concerns the detection of possible functional relations between the tumor samples. Clustered subsets of the tumor samples might give a hint about a specific behavior of the tumor under some conditions. The sample clustering is accomplished in R using the hierarchical clustering.

Single: Distance of the two most similar instances

$$dist(cx,cy) = \min\{dist(a,b) \mid a \in c_x \; b \in c_y\}$$

Complete: Distance of two least similar instances

$$dist(cx,cy) = \max\{dist(a,b) \mid a \in c_x \; b \in c_y\}$$

Average: Average distance

$$dist(cx,cy) = avg\{dist(a,b) \mid a \in c_x \; b \in c_y\}$$

Once the tumor sub-classes are identified, the SNR is calculated for each gene to check how strong does it fit in its tumor sub-class.

III. RESULTS

The application of statistical tests such as the KS test or the t-tests returns a set of differentially expressed genes. Such genes are sent to the wet-lab for testing or are validated based on literature evidence. Recently, Lech et al presented a set of marker genes for the colorectal cancer [14]. We tested the expression of these marker genes in the GEO dataset (GSE32323). Our differential gene analysis using the KS test validated the findings of Lech et al.. Fig. 2 shows the differential expression of the marker gene PIK3CA identified by Lech et al.

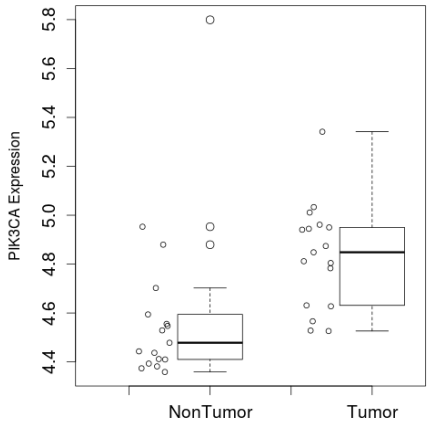

Fig. 2: Marker gene PIK3CA is found differentially to be expressed in GEO dataset GSE32323 (p-value=0.004). this agrees with recent experimental findings.

Next, we searched the whole dataset for all possible correlations setting the PCC threshold to ≥0.9 for correlation and ≤-0.9 for anti-correlation. The expression of the marker gene PIK3CA was found to be correlated or anti-correlated with several other genes. Fig. 3 and table 1 presents one example of each correlation type.

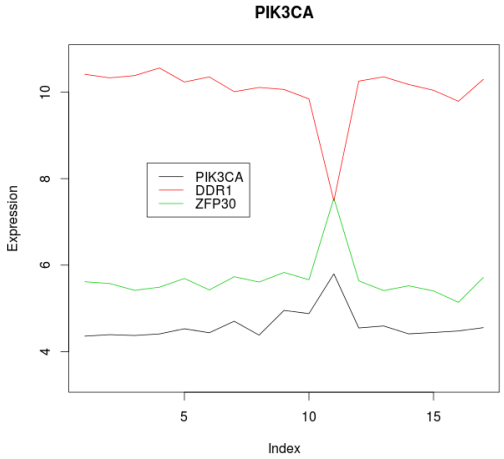

Fig. 3: Plotting one expression correlation and anti-correlation cases with the marker gene PIK3CA. Shown on the x-axis is the expression in different non-tumor samples

Table 1

Gene 1	*Gene 2*	*PCC Tumor*	*PCC Non-Tumor*
PIK3CA	ZFP30	0.1	0.9
PIK3CA	DDR1	0.2	-0.9

Example of PIK3CA correlations and anti-correlations providing the PCC values.

Finally, we grouped tumor samples into sub-classes using hierarchical clustering. We found two major sub-classes within the tumor samples Fig. 4. Class A consists of five samples while class B consists of eleven samples. The SNR was calculated for each gene independently. For example, PIK3CA was placed in class B with SNR value of 9.9. A heatmap of some of class A genes is provided to illustrate that some genes might show clear differential expression even within the tumor samples excluding the non-tumor samples (Fig. 5).

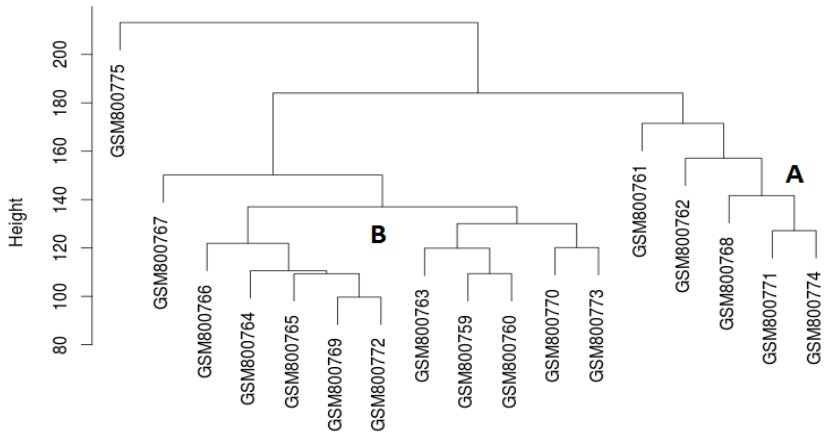

Fig. 4: Hierarchical clustering of the GEO dataset (GSE32323).

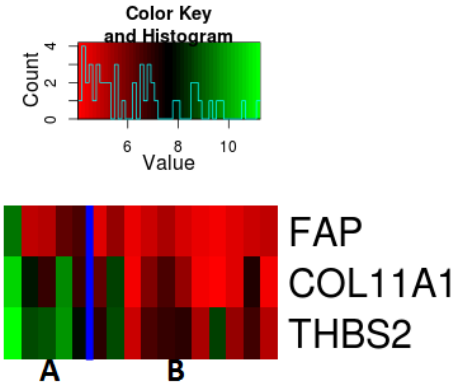

Fig. 5: Heatmap of some class A genes showing a clear differential expression between the two tumor classes

IV. Conclusions

Computational data analysis performed with open source statistical environments may assist cancer research especially in early stages. Such low-cost approaches may be based on publically available large-scale datasets that became available recently what leads substantial savings in time and costs. The statistical environment R-cran incorporates many useful tools and libraries that aid the cancer analysis at the gene and the sample expression levels.

V. References

[1] F. Ducray, J. Honnorat, and J. Lachuer, “DNA microarray technology: principles and applications to the study of neurological disorders,” Revue Neurologique, vol. 163, p. 409–420, 2007.

[2] A. Barghash, N. Golob-Schwarzl, V. Helms, J. Haybaeck, and S. M. Kessler, “Elevated expression of the IGF2 mRNA binding protein 2 (IGF2BP2/IMP2) is linked to short survival and metastasis in esophageal adenocarcinoma,”, Oncotarget, 2016.

[3] A. Barghash, V. Helms, and S. M. Kessler, “Overexpression of IGF2 mRNA-Binding Protein 2 (IMP2/p62) as a Feature of Basal-like Breast Cancer Correlates with Short Survival,” Scandinavian journal of immunology, vol. 82, no. 2, pp. 142-143, 2015.

[4] S. M. Kessler, S. Laggai, A. Barghash, C. S. Schultheiss, E. Lederer, M. Artl, V. Helms, J. Haybaeck, and A. K. Kiemer, “IMP2/p62 induces genomic instability and an aggressive hepatocellular carcinoma phenotype,” Cell death & disease, vol. 6, no. 10, p. e1894, 2015.

[5] A. L. Jackson, S. R. Bartz, J. Schelter, S. V. Kobayashi, J. Burchard, M. Mao, B. Li, G. Cavet, and P. S. Linsley, “Expression profiling reveals off-target gene regulation by RNAi,” Nature biotechnology, vol. 21, pp. 635-637, 2003.

[6] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, “Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data,” Nature genetics, vol. 34, no. 2, pp. 166-176, 2003.

[7] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, “NCBI GEO: archive for functional genomics data sets—update,” Nucleic acids research, vol. 41, no. D1, pp. D991-D995, 2013.

[8] L. Gautier, L. Cope, BM. Bolstad and RA. Irizarry, “affy—analysis of Affymetrix GeneChip data at the

probe level," Bioinformatics, vol. 20, no. 3, pp. 307-315, 2014.

[9] ME. Ritchie, B. Phipson, D. Wu, Y. Hu, CW. Law, W. Shi and GK. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," Nucleic Acids Research, vol. 43, no. 7, p. e47, 2015.

[10] A. Barghash, T. Arslan, and V. Helms, "Robust detection of outlier samples and genes in expression datasets," Journal of Proteomics and Bioinformatics, vol. 9, pp. 38-48, 2016.

[11] V. Carey, "parody: Parametric And Resistant Outlier DYtection," R package, 2016.

[12] Jr. F. E. Harrell, "Hmisc," R Package, 2016.

[13] Y. Hoshida, MB. S. Nijman, M. Kobayashi, J. A. Chan, J. P. Brunet, D.Y. Chiang, A. Villanueva, "Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma," Cancer research, vol. 69, no. 18, pp. 7385-7392, 2009.

[14] G. Lech, R. Słotwiński, M. Słodkowski, and I. W. Krasnodębski, "Colorectal cancer tumour markers and biomarkers: Recent therapeutic advances," World journal of gastroenterology, vol. 22, no. 5, p. 1745, 2016.